High Resolution Multi-View Image-based Building Type Classification Using Deep Learning

Mohammad Hassan Tavakoligargari [1*], Maryam Ghasemzadeh [2], Nima Hazrati [1], Hossein Arefi [1]

[1] i3mainz, Mainz University of Applied Sciences, Germany
[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran,
Email Address: mohamadh.tvkl@gmail.com; maryam.ghasemzad@ut.ac.ir; nima.hazrati@students.hs-mainz.de;hossein.arefi@hs.mainz.de

KEY WORDS: High-Resolution Remote Sensing, Street-Level Imagery, Deep Multi-View Fusion, Stacking Ensemble Learning, CNN, Urban Building Classification

ABSTRACT:

The classification of building types is a major method for optimizing urban planning, enhancing disaster management strategies, and advancing sustainable development objectives. This study presents a multi-view deep learning approach that achieves an overall classification accuracy of 75.8% for distinguishing building types. Using OpenStreetMap (OSM) building tags as ground-truth labels and a multi-view image dataset of 10,360 buildings from the German states of Baden-Württemberg and Rhineland-Palatinate, was generated accordingly. The multi-scale images include aerial images at multiple zoom levels as well as street view images for each building, which are then classified into four categories: commercial, industrial, public, and residential. This approach employs two convolutional neural network architectures (VGG16 and Inception3), with each view trained separately using these CNN model architectures. All CNN models were pretrained on ImageNet before being fine-tuned on the building images. The predictions from the separately trained models were fused using model blending to identify the best combination, followed by a stacking ensemble framework with a Random Forest meta-model for the final classification. Experimental results show that this model fusion leads to a 16% relative improvement in classification accuracy compared to all individually trained models. This paper highlights the importance of integrating different types of views and state-of-the-art CNN architectures, as well as employing model fusion methods for improved urban building classification. Future research will focus on enhancing model fusion techniques and possibly enriching the classification via the incorporation of statistical data on population, income distribution, and infrastructure.

1. INTRODUCTION

The concept of typology is one of the most fundamental and essential ways to classify and arrange social phenomena. Indeed, typology refers to a term, definition, or quality applied to a collection of items, information, or even people who share one or more traits. Type and typology were defined first in 1969 as a phenomenon that exists in all places and times, both in social life and the language of social science (McKinney, 1969). Buildings are the main components of the dynamic world of urban planning, shaping the form and morphology of cities (Wurm et al., 2016). Hence, insight into building typology and its diversity is of particular importance. The perspective on building typology classifies buildings into distinct categories based on their shared attributes (Ballarini et al., 2011). Accessibility to various optical images, ranging from consumer-grade camera phones to complex airborne and satellite platforms, is due to the rapid development of mobile devices, sensor technologies, and social media (Hoffmann et al., 2019). However, each category of imagery has particular challenges and limitations. For example, satellite data generally provides high-quality, high-resolution imagery and enables the identification of spatial features such as size, orientation, shape, and boundaries; it cannot convey detailed visual characteristics of objects, including colour, texture, and pattern (Kang et al., 2018). Street-level imagery, composed of consecutively captured georeferenced photos, provides a valuable data source for evaluating visual traits (Biljecki and Ito, 2021). Nearly all research on building type classification has focused on satellite imagery and remote sensing data. While spatial resolution in this type of image significantly increased, classifying building types using aerial imagery remains challenging, primarily because nadir views provide only rooftop-level information (Zhu and Newsam, 2016). Thus, the necessity for ground-level data seems more obvious and valid. Thankfully, due to the growing availability of ground-level geo-tagged data, it has become easier to fuse remote sensing imagery with data obtained from various sensors and observations (Cao et al., 2018; Lefèvre et al., 2017). Google Street View (GSV) (Anguelov et al., 2010) along with Mapillary, Bing, and other services, offers users panoramic images captured in thousands of cities around the world, allowing them to observe street scenes and offering ground-level information that cannot be obtained from aerial images (Cao et al., 2018). Nowadays, despite the availability of aerial imagery and ground-level data, along with recent developments in computer vision and deep learning, fusing these data remains a difficult task (Hoffmann et al., 2019).

To this end, this article explores the fusion of aerial imagery with street view images using deep learning methods to classify buildings by type into four categories: residential, commercial, public, and industrial. The rest of this paper is organized as follows. In the next section, the background and related works on building type classification will be reviewed, followed by an explanation of the applied methodology. Finally, the results and future work will be discussed.

* Corresponding author

2. RELATED WORK

2.1 Building Type Classification Using Aerial Imagery

Utilizing high-resolution aerial (satellite or drone) imagery for land-cover and land-use mapping has a long historical background. Based on the early research, the employed methods relied on hand-crafted features (spectral, texture, shape) and classical classifiers. Deep learning and learned features transformed remote-sensing classification methods. Fine-tuning OverFeat on the UC Merced land-use dataset achieved 92.4% accuracy, significantly exceeding earlier methodologies (Hoffmann et al., 2019; Sermanet et al., 2014). In contrast, traditional techniques have typically paired hand-crafted color, texture, and geometric features with classifiers such as SVMs, decision trees, or Random Forests (Hoffmann et al., 2019; Xie et al., 2022). Using fully convolutional networks, including U-Net and DeepLab, for semantic segmentation of very high-resolution imagery is another approach that generates accurate building masks and land-cover maps. Finally, CNN-based classification refers to end-to-end classifiers such as ResNet, VGG, and EfficientNet, which are trained on building or roof labels (Hoffmann et al., 2019; Vasavi et al., 2023).

2.2 Building Type Classification Using Ground-Level Imagery

The advantage of street-level imagery over nadir view in building type classification is that the former provides facade details of buildings. One common approach to utilizing these images for building type classification is through deep learning techniques. By fine-tuning multiple convolutional neural networks such as ResNet and VGG on large street-level image datasets like Google Street View, Flickr, and Mapillary, researchers have demonstrated the usefulness of facade structure for identifying building types. Furthermore, coarse-to-fine classification is implemented using hierarchical and multi-label models. Street-level techniques confront challenges such as occlusions, viewpoint variations, and incomplete visualization of buildings with panorama services.

2.3 Building Type Classification Using both Aerial and Ground-Level Imagery

Recent works have focused on fusing aerial and street-level images, as they provide complementary perspectives. Learning the relationship between ground-level photos and overhead imagery was first introduced by (Lin et al., 2013). They presented the cross-view image geolocalization problem in their research. (Workman et al., 2017) proposed a comprehensive model combining ground and remote sensing images, in which features were extracted from sparse ground images and fused with aerial imagery for land use and building function. (Cao et al., 2018) likewise, integrates aerial and street-view data for urban land-use mapping using a SegNet-based architecture. Moreover, (Hoffmann et al., 2019) extended this idea and tried to fuse these complementary views. They evaluated two convolutional neural network strategies: (a) a “two-stream” architecture that fuses feature maps inside the network (geometric-level fusion) and (b) decision-level fusion, which pools the results from two separately trained CNNs. The results revealed that processing aerial and street-view inputs separately and fusing their predictions produced better performance, while combining features early on usually reduced accuracy. Overall, combined strategies include multistream architectures such as two-branch CNNs that process aerial and ground-level images simultaneously, along with late fusion, and multi-resolution imagery(Hoffmann et al., 2019).

3. METHODOLOGY

This section details the dataset preparation, preprocessing, model development, and fusion strategies for building type classification using aerial and street-level images.

3.1 Dataset

A dataset consisting of aerial and street-level images was created for building instance classification. Building footprints for each building were extracted as shapefiles from the Geofabrik download service[2], which offers regularly updated OSM extracts. Using these shapefiles, the centroid geometries of the buildings were calculated. Then, each building's functionality was extracted through OSM, and this data was used to group buildings into 16 categories, which were then simplified into the four main building types (Table 1): commercial, industrial, residential, and public. Three zoom levels (19, 20, 21) were chosen to capture various levels of information. A lower zoom level offers a broader neighbourhood perspective, while a higher one reveals more details, such as roof materials. This multi-scale approach allows a fusion network to prioritize detailed information for each building dynamically. Based on the geometric data obtained for each building, a custom Python script was used to automatically retrieve the corresponding nadir view image via the Google Map Tile servers, while the Google Street View image was acquired by querying the Google Street View API. The aerial images were adjusted so that the target building approximately appears at the center of the image, aligned with its geolocation. The reasoning behind this is to ensure consistent image framing by removing translation variance, to make precise alignment with street view images through spatial correspondence, and to maximize the utility of the image area by centering the buildings and avoiding irrelevant background objects. Street View images were queried via the Google Street View API with a focus on comprehensively capturing building facades. The pitch was set to 0° and the field of view (FOV) to 90° to ensure consistency. To calculate the heading value, the coordinates of the location where the Street View image was taken must first be determined. This information is available in the metadata of each image retrieved via the Google Street View Static API. These coordinates, together with the coordinates of the building center, define a line. The heading value can then be calculated using Equation (1). Point 1 represents the location where the Street View image was captured and Point 2 represents the centre of the building.

$$Heading = mod\ (atan2(\sin(\Delta\lambda).\cos(\varphi_2),\ \cos(\varphi_1).\sin(\varphi_2) - \sin(\varphi_1).\cos(\varphi_2).\cos(\Delta\lambda) \times 180/\pi + 360,360) \quad (1)$$

where φ_1, φ_2 = the geographic latitudes of points 1 and 2 in radians
λ_1, λ_2 = their corresponding longitudes, in radians
$\Delta\lambda$ = the difference in longitude of two points

The images cover two German federal states: Rhineland-Palatinate and Baden-Württemberg. The dataset includes four images for each building location: three aerial images captured at three different zoom levels, and one corresponding street-level view image. Both the aerial and street-level photos were retrieved from Google Maps and Google Street View in 2024.

[2] https://download.geofabrik.de

As shown in Table 1, Raw tags were aggregated into target classes, resulting in a total of 10,361 building instances, each associated with four images. Also, all photos were resized to 512×512 px. Moreover, it is essential to note that this study employs a supervised learning approach to train the CNN models. This means the images must be labeled according to their respective categories. Further details regarding this are provided in the following section.

Classes	OSM Tags	Number of Buildings
Commercial	Commercial	2559
	Retail	
	Office	
Industrial	Industrial	1211
	Warehouse	
Public	Church	3588
	College	
	Hospital	
	Hotel	
	Public	
	School	
	University	
Residential	Apartment	3003
	Dormitory	
	House	
	Residential	
In Total		10,361

Table 1. Building Tags and Instance Numbers

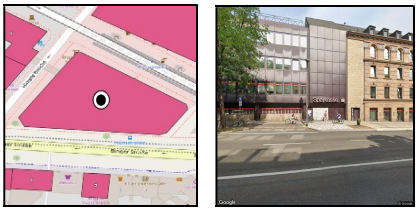

Figure 1. Left: Centroid of Building; Right: Google Street View Image

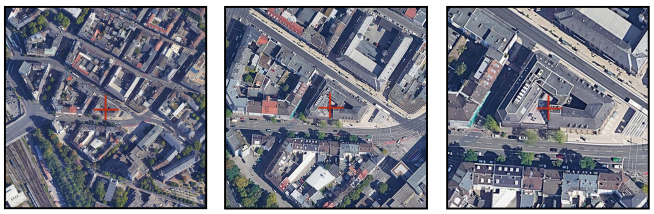

Figure 2. Left: Aerial Image Zoom 19; Middle: Aerial Image Zoom 20; Right: Aerial Image Zoom 21.

3.2 Model Development

The VGG16 and Inception3 architectures were chosen for their complementary capabilities in training with limited remote sensing data and their greater compatibility with the building images. According to the work (Hoffmann et al., 2019) VGG16 offers a uniform, compact design that makes it easy to fine-tune and interpret on limited data. By comparison, Inception3 maintains a modest computational cost while achieving high accuracy by using factorized, multi-scale convolutions. Various architectural designs, e.g., the deep residual blocks used in ResNet, add extra complexity that can cause overfitting with limited data. Likewise, the complex scaling method used in EfficientNet requires extensive hyperparameter search and tuning, which could bring limited gain for comparatively small remote sensing datasets.

To tackle the task of building type classification, and given the small size of the classification dataset, transfer learning was employed by pre-training VGG16 and InceptionV3 on the ImageNet dataset to leverage the rich visual features learned from large-scale datasets. ImageNet provides robust object-level representations that are particularly useful for both aerial and street-view imagery. Fine-tuning was then performed on the image dataset of each view. In this process, the networks were initialized with ImageNet-pretrained weights to leverage generic visual features such as edges, shapes, and textures learned from large-scale natural images. The early convolutional layers, which capture these low-level and transferable features, were frozen to retain their pretrained representations, while the later layers—responsible for learning higher-level, domain-specific patterns—were unfrozen and fine-tuned using the labeled building-type dataset. This strategy allows the models to adapt effectively to the characteristics of aerial and street-level imagery while avoiding overfitting due to limited domain-specific data. The pre-training and fine-tuning procedures were conducted separately for each view, resulting in a dedicated classification model per view. Due to the imbalanced dataset, we used class-weighted loss functions.

After pre-training and fine-tuning each model, to improve accuracy and achieve a unique classification, a model fusion approach comprising model blending and model stacking techniques was applied.

As shown in Figure 3, we define model blending as a weighted averaging of prediction vectors from base models. In model blending, each separately trained model generates predictions on the validation set, resulting in a prediction vector. These predictions are then combined using a weighted average, where each model is assigned a specific weight. This produces a merged prediction, which is considered the final output of the ensemble model. In the case of a classification task, this final prediction represents the probability distribution over the classes. The goal of model blending is to examine how classification accuracy changes after integrating the models, to find the best model combination.

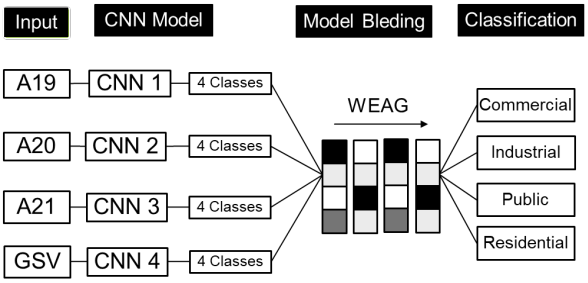

Figure 3. The Procedure for Model Blending

As shown in Figure 4, Model stacking is different from model blending by employing a multi-level training process. First, the predictions of separately trained models on a validation dataset are gathered as probability vectors, serving as input features for

a higher-level model known as the meta-model. This meta-model often uses a simple classification algorithm such as logistic regression, a decision tree, or a random forest, and is trained to combine the base models' outputs to produce the final prediction.

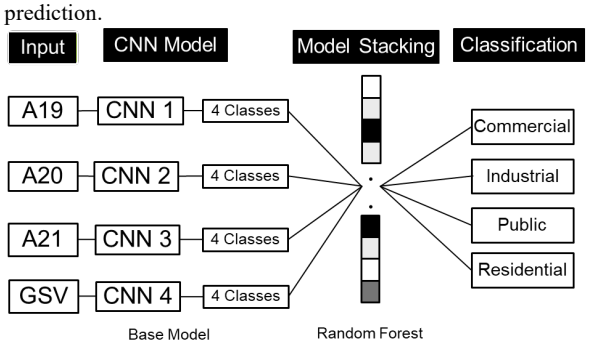

Figure 4. The Procedure for Model Stacking

More details regarding the training parameters, implementation, and evaluation are presented in the following section.

4. EXPERIMENTAL SETUP AND EVALUATION

The presented method, which involves training the models for each view separately using VGG16 and Inception3, and applying model fusion techniques for the final classification, is implemented in this section for both federal states in Germany: Rhineland-Palatinate and Baden-Württemberg. The entire implementation process is illustrated in Figure 5. In this study, decision-level fusion was employed, consisting of model blending and model stacking approaches for integrating the outputs of the separately trained CNN models and obtaining the final classification results. During implementation, we'd periodically peek at intermediate results; sometimes our fused models surprised us with odd misclassification. Once each fusion step wrapped up, we ran a final evaluation to see exactly how much our ensemble boosted accuracy.

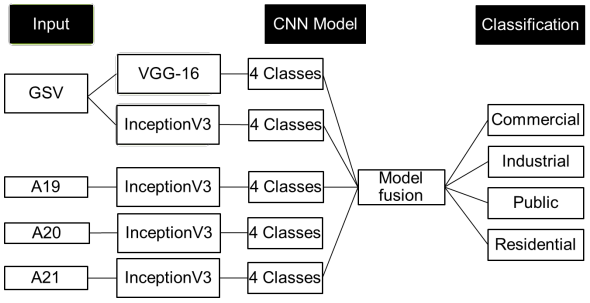

Figure 5. The Procedure for The Experimental Setup

4.1 Separate Training of Each CNN Model

Selecting an appropriate CNN model and choosing the right training hyperparameters play crucial roles. These parameters vary depending on the dataset and the desired target accuracy. After testing both CNN models on each dataset and their respective hyperparameter configurations and evaluating intermediate results, the final parameters for training each CNN model were determined. InceptionV3 was selected for both aerial and street-level images due to its ability to capture multi-scale spatial patterns and extract both local and global contextual features effectively. Its inception modules are particularly suitable for handling variations in scale, texture, and illumination that are common in aerial imagery and street-level scenes. VGG16 was also employed to provide a complementary baseline with a simpler and more uniform architecture, enabling comparative evaluation and feature diversity in the model fusion process. Table 2 summarizes the final training settings for each model and dataset.

Image Type	Model	Batch Size	Dropout Rate	N	L
Aerial (19-21)	Inception3	32	0.2	20	2×10^{-4}
				10	1×10^{-4}
Street View	VGG16	64	0.2	20	2×10^{-4}
				10	1×10^{-4}
	Inception3		0.35	10	5×10^{-5}
				10	2×10^{-4}
				10	1×10^{-4}

Table 2. Training Parameters (N=epochs, l= Learning rate)

Based on the determined model architectures and parameters, the process of pre-training and fine-tuning the CNN models was initiated. As mentioned in the dataset section, the images were labeled according to their respective categories. The dataset was then divided into three subsets: training, validation, and test. The training set consists of 75% of the entire building dataset (1,295 images per view), a choice based on our need for robust learning curves, while the validation and test set each account for 12.5% (1,295 images per view, respectively). Each view is fine-tuned separately on its corresponding training dataset using the defined model architecture and specified hyperparameters. We've found these yields consistently stable losses. Subsequently, all trained models are evaluated on the test dataset using the confusion matrix and performance metrics such as precision, recall, Cohen's kappa, F1 score, and accuracy. The calculated metrics for each model are presented in Table 3, and the confusion matrices are shown in Figures 6 to 9, results that, in our experience, offer a solid roadmap for further tuning. In aerial images, Inception3-A19 shows the lowest accuracy, while Inception3-A21, with its high recall and kappa, provides more balanced and reliable classifications and achieves the best accuracy. Models based on street view images also deliver strong results, with Inception3-GSV overall performing better than VGG16-GSV.

Model	Prec. (%)	Rec. (%)	F1	Kappa	Acc. (%)
Inception3 - A19	57.2	56.8	57.2	0.47	60.0
Inception3 - A20	76.9	66.2	69.2	0.55	66.2
Inception3 - A21	73.3	70.0	60.1	0.60	69.7
VGG16-GSV	65.9	63.1	66	0.51	63.1
Inception3 - GSV	71.7	68.5	70	0.58	68.5

Table 3. Evaluation After Separate Training of The Models

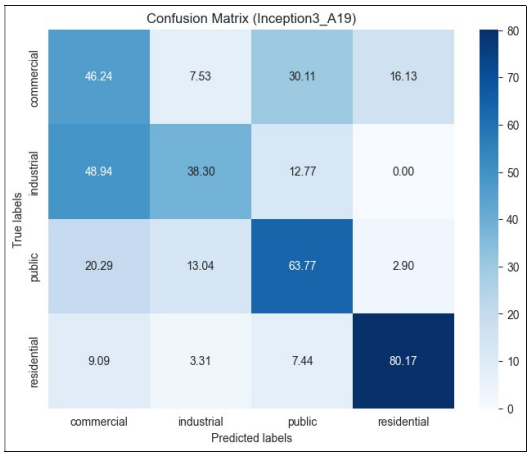

Figure 6. Confusion Matrix of Inception3- A19

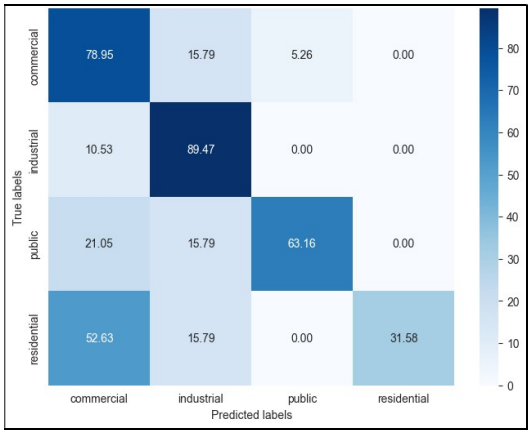

Figure 7. Confusion Matrix of Inception3- A20

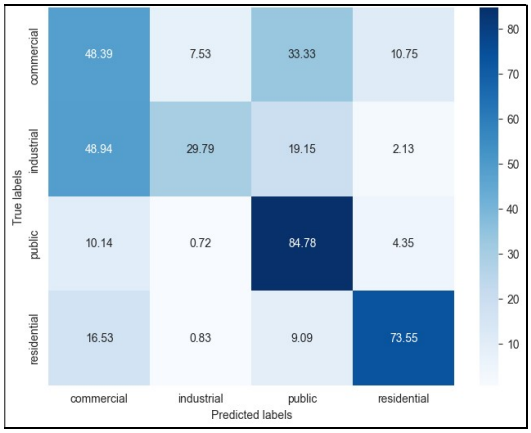

Figure 8. Confusion Matrix of Inception3- A21

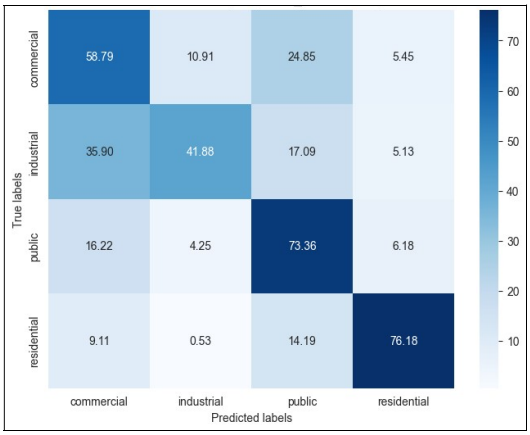

Figure 9. Confusion Matrix of VGG16- GSV

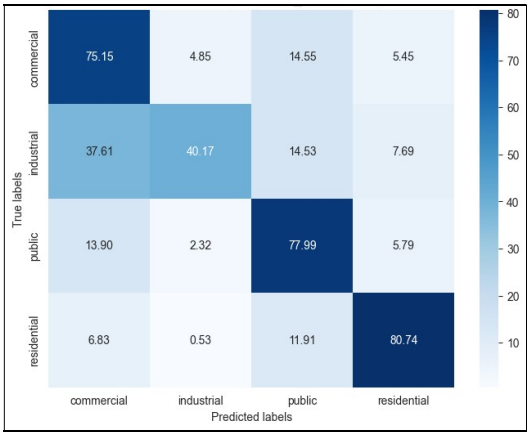

Figure 10. Confusion Matrix of Inception3-GSV

4.2 Model Blending and Model Stacking

Since this study used three different scales of aerial imagery, it is important to select the best zoom levels for model fusion. Therefore, examining the change in accuracy after model fusion through model blending for aerial images is necessary. To this end, the trained models from aerial images at zoom levels 19 (Inception3-A19) and 20 (Inception3-A20) will be combined for Fusion 1, the trained models at zoom levels 21 (Inception3-A21) and 20 for Fusion 2, and the models at zoom levels 19 and 21 for Fusion 3. The results of model blending are presented in Table 4, showing that Fusion 1 achieved an accuracy of 58%, Fusion 2 achieved 62%, and Fusion 3 achieved 65%.

Fusion	Model	Accuracy	
1	Inception3-A19	60.0 %	58.3 %
	Inception3-A20	66.2 %	
2	Inception3-A21	69.7 %	62.0 %
	Inception3-A20	66.2 %	
3	Inception3-A19	60.0 %	65.1 %
	Inception3-A21	69.7 %	

Table 4. Evaluation of Model Blending for Aerial Images

It is apparent that integrating the models trained on aerial images from all zoom levels with zoom level 20 results in lower accuracy after integration. After fusion with zoom level 20, the

overall accuracy is lower than that of each model. This is because zoom level 20 does not provide additional information compared to other zoom levels, such as 19 and 21, which causes the model to be more prone to overfitting after integration. For this reason, removing the model trained on images from zoom level 20 from the list was decided. In the end, four separately trained models remain: two models trained on aerial images from zoom levels 19 and 21 (Inception3-A19 and Inception3-A21), and two models trained on street view images (Inception3-GSV and VGG16-GSV). These four models are then used for integration. After evaluating the accuracy of the model combinations, the model with the best performance for fusion was identified, the final fusion approach is implemented using model stacking, where the base model consists of four separately trained models on aerial (Inception3-A19, Inception3-A21) and street view images (Inception3-GSV, VGG16-GSV), while the Random Forest classification method is selected as the meta-model. First, each base model generates a prediction (i.e., probability) using their respective validation set. Second, the predictions are aggregated so that each image is described as a set of probabilities, with every probability coming from a different model. These combined predictions are the input features for the training of the meta-model, which learns to determine the best way to combine the outputs of the base models to produce the final classification decision.

4.3 Evaluation

Once the meta-model is trained, the base models produce predictions on the test data, which are combined and fed into the meta-model to generate the final classification result. To evaluate the model fusion using model stacking, the confusion matrix and metrics such as precision, recall, kappa, F1 score, and accuracy are used to determine the classification performance for the correct classification of building types in the dataset. The confusion matrix is presented in Figure 11, and the evaluation metrics are summarized in Table 5.

Building Type	Precision	Recall	F1	Kappa	Accuracy
Commercial	68.7%	75.2%	71.9%		
Industrial	90.0%	69.1%	78.0%		
Public	68.3%	78.0%	72.8%	0.67	75.8%
Residential	81.0%	80.7%	80.8%		
WAVG	77.5%	75.7%	76.6%		

Table 5. Classification Performance after Model Fusion

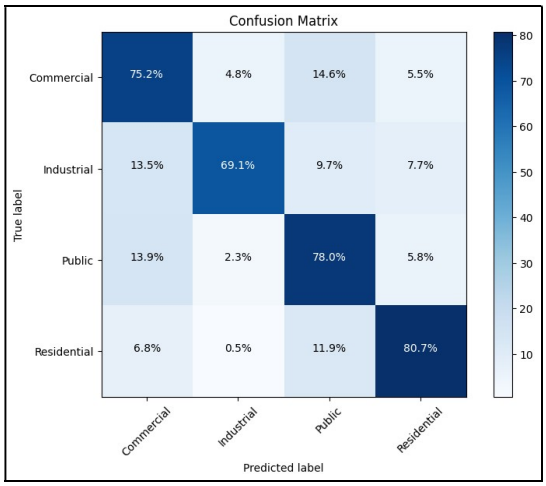

Figure 11. Confusion Matrix after Model Fusion

5. RESULTS AND DISCUSSION

An analysis of the confusion matrix (Figure 11) for the model fusion approach described in the previous section indicates that the classification model exhibits strong overall performance. For a four-class classification problem, the model much surpasses the random baseline with an overall accuracy of about 75.8%. Moreover, Cohen's Kappa value of about 0.67 (Table 5) shows a rather high degree of agreement between the expected and actual class labels. With proper classification rates of 80.7% and 78.0%, respectively, the model performs especially well in spotting the "Residential" and "Public" classes. This implies that these groups have unique and readily identifiable characteristics. Performance is comparatively lower for the "Industrial" class, sometimes misclassified as "Commercial" (13.5%) or "Public" (9.7%). Analogously, in 14.6% of the cases, the "Commercial" class is misclassified as "Public". These incorrect classifications demonstrate that Industrial and Commercial categories are harder to distinguish due to shared architectural or façade-level traits such as large structures, signage, or similar material textures. Table 5 provides a summary of the key performance metrics. Precision and recall values, which usually range from 75% to 78%, indicate a strong overall classification performance. It's interesting to note that the "Industrial" class has a high precision of about 90%, meaning that it is typically accurate when predicted. Its low recall of roughly 69%, however, indicates that many real class instances are missing (i.e., false negatives). The efficiency of the model fusion approach is shown by comparing Tables 3 and 5. The fusion of models improved the overall accuracy by up to 16% compared to the best-performing individual model.

Figure 12 shows examples of correct predictions in green and incorrect predictions in red. A closer inspection of misclassified instances reveals that a significant portion, approximately 60% of industrial to commercial errors, occur in buildings with large, rectangular facades, visible loading bays, or signage – features commonly found in both categories. Similarly, commercial to public misclassifications often involve large institutional-looking buildings, suggesting a shared visual vocabulary between modern civic and retail architectures. These observations highlight that while the model captures general visual cues effectively, class boundaries with overlapping morphological characteristics remain the main source of confusion.

6. CONCLUSION AND FUTURE WORK

This study demonstrates the potential of deep learning models, combined with OpenStreetMap data and multi-view imagery, and using model fusion methods to accurately and reliably classify building types in German urban areas.
For future work, an early fusion approach is planned instead of training separate models and applying model fusion at the decision level. All available data sources—including additional building attributes such as material and area—will be integrated at the input level and used to train a single CNN model, in order to save training time and achieve higher accuracy across all categories. Moreover, once building types have been classified, they can be enriched with statistical data on population, income distribution, or infrastructure, thereby enabling deeper insights into social diversity, economic activity, and transportation networks within urban neighborhoods.

Figure 12. Examples of Correct (Green) and Wrong Predictions (Red)

REFERENCES

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google Street View: Capturing the World at Street Level. Computer 43, 32–38. https://doi.org/10.1109/MC.2010.170

Ballarini, I., Corgnati, S.P., Corrado, V., Talà, N., 2011. DEFINITION OF BUILDING TYPOLOGIES FOR ENERGY INVESTIGATIONS ON RESIDENTIAL SECTOR BY TABULA IEE-PROJECT: APPLICATION TO ITALIAN CASE STUDIES.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning 215, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating Aerial and Street View Images for Urban Land Use Classification. Remote Sensing 10, 1553. https://doi.org/10.3390/rs10101553

Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model Fusion for Building Type Classification from Aerial and Street View Images. Remote Sensing 11, 1259. https://doi.org/10.3390/rs11111259

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building Instance Classification Using Street View Images. ISPRS Journal of Photogrammetry and Remote Sensing 145, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006

Lefèvre, S., Tuia, D., Wegner, J.D., Produit, T., Nassar, A.S., 2017. Toward Seamless Multiview Scene Analysis From Satellite to Street Level. Proceedings of the IEEE 105, 1884–1899. https://doi.org/10.1109/JPROC.2017.2684300

Lin, T.-Y., Belongie, S., Hays, J., 2013. Cross-View Image Geolocalization, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, OR, USA, pp. 891–898. https://doi.org/10.1109/CVPR.2013.120

McKinney, J.C., 1969. Typification, Typologies, and Sociological Theory*. Social Forces 48, 1–12. https://doi.org/10.1093/sf/48.1.1

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. https://doi.org/10.48550/arXiv.1312.6229

Vasavi, S., Sri Somagani, H., Sai, Y., 2023. Classification of buildings from VHR satellite images using ensemble of U-Net and ResNet. The Egyptian Journal of Remote Sensing and Space Sciences 26, 937–953. https://doi.org/10.1016/j.ejrs.2023.11.008

Workman, S., Zhai, M., Crandall, D.J., Jacobs, N., 2017. A Unified Model for Near and Remote Sensing. Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, pp. 2707–2716. https://doi.org/10.1109/ICCV.2017.293

Wurm, M., Schmitt, A., Taubenböck, H., 2016. Building Types' Classification Using Shape-Based Features and Linear Discriminant Functions. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9, 1901–1912. https://doi.org/10.1109/JSTARS.2015.2465131

Xie, X., Liu, Y., Xu, Y., He, Z., Chen, X., Zheng, X., Xie, Z., 2022. Building Function Recognition Using the Semi-Supervised Classification. Applied Sciences 12, 9900. https://doi.org/10.3390/app12199900

Zhu, Y., Newsam, S., 2016. Land Use Classification using Convolutional Neural Networks Applied to Ground-Level Images. https://doi.org/10.48550/arXiv.1609.06653